国际视野下治理虚假新闻的技术手段及相关模型

李 净
（《中国传媒科技》杂志社，北京 100031）

摘 要：伴随着人工智能、社交媒体、大数据、VR/AR 等技术的进一步发展，假新闻混入信息的汪洋大海，无孔不入，难以辨别，引发了全球的传播危机、社会危机，已经成为一个不容忽视的议题。本文通过梳理目前治理虚假新闻的技术手段，总结国际经验，以求为国内打击虚假新闻、净化媒体环境、营造清朗的媒体空间提供参考和借鉴。

关键词：假新闻治理；事实核查；人工智能；假新闻检测技术；深度伪造　　**中图分类号**：G210　　**文献标识码**：A
文章编号：1671-0134（2021）08-017-05　　DOI：10.19483/j.cnki.11-4653/n.2021.08.003
本文著录格式：李净．国际视野下治理虚假新闻的技术手段及相关模型 [J]. 中国传媒科技，2021（08）：17-21.

导语

假新闻由来已久，但由于近年来社交媒体的低门槛和广泛应用，以及深度伪造等新技术的加持，社会原有的甄别及防范机制难以及时做出反应，假新闻掀起一波又一波浪潮，逐渐泛滥成灾，严重冲击了全球的政治、经济秩序，成为人类公害。在中国知网平台上以“假新闻”为主题搜索相关论文，可以看到目前国内对假新闻的研究主要集中在网络舆论和政治传播等维度，关注假新闻生成原因、扩散机制、社会危害、监管治理等方面。其中治理对策偏重于法律监管和行业自律，技术探讨相对较少。对比同时期相关的外文文献，发现其更关注具体的打击假新闻的技术手段。相对于法律法规、道德约束、媒介素养等需多方联动、长久发力的治理措施，当下以技术为主导参与假新闻治理更加实际、可操作性更强。本文通过分析目前打击假新闻的常用方法以及主要检测技术，探索人工智能参与假新闻治理的国际经验，以求为国内打击虚假新闻、净化媒体环境提供参考和借鉴。

1. 早期治理假新闻的常用方法

早期治理假新闻的常用方法包括：事实核查、来源验证、评论核查、新闻内容特征核查、用户参与度以及个人资料与文章、读者、发布者之间的关系核查等。[1] 其中，事实核查和来源验证是打击假新闻最普遍的做法。

英国“第四新闻频道”在 2010 年就创建了打击假新闻部门 “事实核查”（FactCheck）；2011 年，《卫报》推出事实核查博客网站 Reality Check；2015 年，BBC 创建事实核查团队“现实核查”；2016 年事实检查机构 Full Fact 推出自己的监控系统，用于掌握每个谣言的生命周期。目前，打击假新闻的普遍做法是在核查后的信息上打标签。譬如 Facebook 和 Twitter 等社交平台和以 Google News 为代表的搜索引擎对可疑内容标记上“正在由第三方机构进行事实核查”，这些内容不被平台优先推荐，处于新闻流的末端。Facebook 还设置了 24 小时举报功能，以尽可能早地发现不实信息。《世界报》推出“Dé codex”事实核查数据库，利用颜色编码系统，让读者了解某一网站的可靠性，如绿色代表高度可靠，黄色代表谨慎阅读，红色则意味着该网站会发布虚假信息或完全编造的故事。讽刺性网站标记为蓝色，而一些不能被验证的网站则被标记为灰色。[2] 意大利在 2012 年创建了事实核查网站 Pagella Politica。这些核查工作有力地遏制了假新闻的传播，但因信息数量的急速增长，核查需要更多的技术支撑。

2. 近年来假新闻治理的研究方向

有统计表明，目前散布假新闻的账户中，三分之一到三分之二是社交机器人账户。[3] 假新闻在生产传播过程中运用的主要技术包括算法技术（Algorithmic）、深度伪造技术（Deepfake）、机器学习技术（Machine learning）等。其中，算法推荐指引着流量的流向，用户分析实现精准化与智能化，平台在“流量思维”驱使下，造假、造谣屡禁不止。而深度伪造技术和机器深度学习、自动决策技术让信息实现了图像与视频的自动生产，新闻生产分发日趋机器化、智能化，加大了信息的可操纵性。随着人工智能变得越来越复杂，机器人制作、分发的新闻成了假新闻的重灾区，捉襟见肘的人工核查已经不能有效打击海量且日益隐蔽的假新闻，打击假新闻的研究目光转向了自动化技术。

《世界报》的事实核查团队就一直在与数据科学家合作，探索如何利用自动化技术及时发现假新闻。2019 年华盛顿大学和艾伦人工智能研究所的研究者提出了一种名为 Grover 的鉴定模型。它从 Google 新闻上 5000 个不同媒体撰写的新闻中进行学习，在此期间接收了 120 千兆字节的真实新闻文章，并用这些文章进行训练，最终 Grover 分辨人和 AI 写的故事的正确率是 92%。在此之

前，最好的假新闻识别器正确率是 73%。Grover 之所以能如此有效地发现虚假内容，是因为它也非常善于自己制作内容。以彼之道还施彼身，用技术抵御“技术‘制造’”，是目前我们治理假新闻的最佳手段。

3. 假新闻治理的技术路径

治理假新闻的具体技术路径在于如何更好地掌握规律、制定规则、确立模型，给算法提供正确指引，让人工智能能够更有效地甄别虚假和低俗内容并控制其传播。[4] 这方面虚假新闻检测技术跑在最前面。虚假信息检测技术是指通过自然语言处理、社交挖掘、跨模态分析等智能处理方法，发现并利用信息的内在特征、产生机理与传播规律，识别并干预假新闻的传播。其中运用的技术包括传统的机器学习算法（比如逻辑回归算法、支持向量机和随机森林等算法）、深度学习（包括、卷积和递归神经网络）和其他模型（矩阵分解和贝叶斯推理等模型）。[5] 下文从新闻内容特征、预测性检测、可解释性检测等方面选取最新的检测模型，逐一介绍。

3.1 基于内容特征的虚假新闻检测技术

根据内容特征的不同，可分为基于文本的、基于图像的，以及基于多模态的虚假信息检测。

3.1.1 以文本为对象的检测

基于文本的检测是虚假信息检测最常用的方法。大多数研究利用文本内容和传播过程中产生的社交上下文（基于用户行为可信度或基于传播网络或语义和情感度量）、[6] 根据虚假新闻特定的语言风格，如论文《Capturing the Style of Fake News》利用文档的风格，对不同来源的数据分类，为新闻的可信度进行打分，从而检测假新闻。早期采用提取语言学特征、主题特征等特征的方法，[7–9] 近年来使用深度模型自动学习数据高层特征的方法，[10] 基于社交上下文的方法主要包括基于用户行为可信度的方法，[11] 以及基于传播网络的方法。[12–15] 论文“基于主题与情感联合预训练的虚假评论检测方法”抽取评论的语义和情感上下文特征，并进行联合训练和优化。[16]

3.1.2 以图像为对象的检测

面向图的异常检测工作最早发表于 2003 年，[17] 图异常检测是指在一个大图或海量图数据库中寻找包含陌生或异常模式的结构（包括节点、边或者子图），研究对象可分为静态图和动态图两类 。静态图上的异常通常是指图中有明显偏差的节点、边或子图。动态图则因为会随着时间变化，产生新的节点或边的增加和删除，所以异常通常是导致变化或事件发生的 topk 个节点、边或子图。

图异常检测通常基于图的结构信息检测（包括节点与节点之间、节点与子图之间，以及子图与子图之间的异常）；[18–20] 基于子空间选择，检测在节点特征的子空间中的异常，[21–23] 以及基于概率统计等进行统计学分析，[24–27] 获取图的统计信息进行检测；动态图经常是先获取图的概要，然后通过聚类和异常检测来确定概要中的异常，例如文献。[28–29] 同时利用深度学习方法检测图像也正在被不断探索研究。[30]

目前在业界，针对图像的检测，微软运用了 Face X-ray 技术，提出通用的检测不同模型生成的合成图的方法，核心是去学习换脸的边界，方法泛化性能优良。芝加哥大学的 Fawkes 技术，可为私人照片提供人眼不可见的像素级保护，避免用户被未知第三方人脸识别模型检测并追踪。脸书推出名为“Rosetta”的 AI 系统，用于帮助计算机理解并分析每天发布在平台上海量的图像与视频。

3.1.3 以音频为对象的检测

语音伪造常见方式为：语音模仿、录音重放、语音合成和语音转换，[31] 检测主要运用的技术是卷积神经网及其变种。工作方向集中在以下几个方面。

（1）利用音频的声学特征。选取可区别真实语音和伪造语音的声学特征。如 Todisco 等人将常数 Q 变换倒谱系数（constantQcepstralcoefficients，CQCC） 应用于语音鉴伪中，Sahidullah 等人用线性滤波代替了梅尔刻度滤波，提出了线性频率倒谱系数（linearfrequencycepstrumcoeffi G cients，LFCC），更关注高频段特征。

（2）设计可以学习到真伪语音的区分表示的分类模型。高斯混合模型（Gaussianmixturemodel，GMM）是之前最常用的分类模型。随着深度学习技术的进展，卷积神经网络（convolutionneuralnetwork，CNN）的性能比直接使用 GMM 更好。例如具有最大特征图（maxfeaturemap，MFM）激活功能的轻量型卷积神经网络（lightconvolutionneuralnetwork，LCNN），通过竞争学习的方法，不只能分离噪声信号和信息信号，还能起到特征选择的作用。以上两种方法都被证明有效，说明使用适当的前端声学特征和深度学习模型都很重要。[32]

3.1.4 基于多模态的检测

多模态（multimodality）的虚假新闻检测难度比较大，也是当前研究的热点方向。文本、图像和视频等共同构成多模态信息，彼此互相支持，辅助证明，增加了可信度也增强了辨别真假的难度。譬如使用 Deepfakes 换脸，又譬如之前德美研究人员研发的一项实时运动捕捉技术，可以将任何演员的面部表情转换成特朗普和普京等政治人物的视频片段、自动替换视频中人的发言内容。英伟达的新版 StyleGAN，图像部分属性（Style）实现解耦的能力催生了大量利用其进行图像编辑的工作，如非常火爆的图像创作工具 Artbreeder （https：//www.artbreeder.com）；香港科大的 InterFaceGAN，提出潜在空间结构

GAN 生成空间的方法，可泛化迁移到所有 GAN 生成的各种人脸样本空间，包括属性编辑、风格转换等。这些新技术的采用加大了多模态假新闻核查的难度。

目前基于多模态的检测方法一般采用通用的循环神经网络（recurrentneuralnetwork，RNN）和卷积神经网络（convolutionalneuralnetwork，CNN）分别捕捉虚假新闻文本及视觉模态表现层面的特性。[33]

具体的检测模型研究有：通过使用文本和视觉特征提取器，并联合两类特征训练的检测器；使用可变自编码器学习文本和图像的共享表示，从共享的潜在表示中分别重构文本和图像，从而捕获两种形式之间的关联性；分别学习文本和图像表示后，计算文本和视觉表示之间的相似性，识别模式之间的"不匹配"；又或是在模型中加入 3 种类型的模块（文本、图像和用户个人资料），结合用户评论的情感，判断文章 / 帖子真假；还有利用外部知识图谱学习基础知识，再结合文本和视觉特征进行检测等，不一而足。[34] 在业界，成立于 2017 年的旧金山人工智能基金会 AI Foundation 开发的 Reality Defender 系统，借助对图像、视频和其他媒体内容的扫描，利用人工智能驱动的分析技术帮助人们识别由人工智能算法生成的内容，以检测潜在的假新闻。

3.2 可解释性检测

可解释性检测是虚假新闻检测研究中的新领域。公众质疑某些信息是否是出于某种政治或经济目的被打上可疑或虚假标签，为什么被打上某种标签？依据是什么？因此解释一篇文章为什么被打上某标签也成为研究新方向。论文《dEFEND：Explainable Fake News Detection》提出了具有可解释性的假新闻检测模型 dEFEND，在社交媒体上的假新闻检测领域是第一个尝试提出具有可解释的模型的研究。此模型分别编码新闻内容和用户评论组件，并利用层级注意力机制和共同注意力机制（sentence-comment co-attention subnetwork）捕获内容和评论之间的关联，最终捕获了可解释的 top-k 个值得检查的句子和用户评论，模型检测效果良好。其他论文如《传播 2Vec：嵌入部分传播网络进行可解释假新闻的早期检测》等模型也进行了相关工作，研究空间依然广阔。

3.3 可预测性检测

谣言的传播模式往往与真实信息的传播非常不同。实现谣言可预测能达到事半功倍的效果。Soroush 等人分析 Twitter 平台的谣言传播模式发现，与真实消息相比，谣言传播的影响范围更广，真实消息在任意一个层级上参与转发的最多人数达到 1000 以上，而谣言的最大转发数最多可达万级。辟谣的速度远远不及传播的速度。因此从根源上、在早期就阻断假新闻的传播，也是目前的一个研究方向。论文《Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks》新近提出了一个适用于"谣言的早期检测"的"Bi-GCN"模型，和以往方法不同的是，该模型考虑到了"自顶而下"的谣言传播（propagation）结构，和"自底而上"的来自不同社区的谣言散布（dispersion）结构。同时使用到了"根源帖子特征的增强"。具体来说就是在 GCN 每层 GCL 中，对每个节点，将根源帖子在上一层的隐层特征表示和节点在该层的隐层特征表示向拼接起来，作为节点在该层的最终隐层特征表示。这种方法增强了谣言根源帖子对学习到其他帖子节点表示的影响力，可帮助模型学习得到更有助于谣言检测的节点表示。这也是"第一个"使用"基于 GCN 的方法"进行谣言检测任务的模型。

另外，基于传播结构的检测方法也是热点研究方向。英国科技公司 Fabula AI 利用几何深度学习的方法，着眼于信息如何在社交网络上传播以及谁在传播，根据真假新闻的可信程度对内容进行分类，给出评定分数。能够用更快的速度，在内容发布后的短时间内，以非常准确的方式检测出假新闻。再者，目前人工智能打击检测假新闻是主流，但用区块链技术，如 Userfeeds 与 PressCoin 模式也是一个值得探讨的方向。[35]

4. 现有的假新闻治理困境

综观目前治理假新闻的技术发展现状，可以看到运用人工智能技术搭建模型检测虚假新闻取得了很大进展。然而，现有研究大多没有使用真实的社交网络上的信息传播数据。模型需要的数据集相对缺乏，尽管有些研究是基于真实的社交网络拓扑结构，但是具体的信息传播与抑制过程完全使用仿真的方式，缺乏真实性。这也造成了学术研究模型停留在思路上，仅限发表在论文中，而没有与实践相挂钩。希望有志于此的人，不管来自媒体还是科技企业、投资公司，加大对优秀论文的关注度，力求将学术成果转化为业界实践，发挥更大作用。

5. 未来治理假新闻可能的方向及趋势

任何信息的真相和意图都不能只靠计算机评估，用技术手段虽然在短期内有效，但治标不治本，属权宜之计。假新闻必然会随技术的发展与人类长期共存，若想实现对假新闻的长期有效治理，仍然需要依靠人工和技术之间的合作，同时从法律法规的制约和公民的道德培养、媒介素养上入手，才是治理假新闻的终极途径。特别是在提升公民媒体素养方面，小—中—大学的教育要承担起帮助青少年培养识别假的新闻能力，开展相关的媒介素养课程，培养批判性思维，大学还应将新闻查证和事实核查作为基础学科，从信息核查能力的技能式培养转向提高自我认知，突破认知偏见，最终实现"谣言止于智者"。治理假新闻，对所有人来说，依旧任重而道远。媒

参考文献

[1] Amrita，Bhattacharjee，舒凯，高旻，刘欢 . 网络信息生态系统中的虚假信息：检测、缓解与挑战 [J]. 计算机研究与发展，2021（7）：1353–1365.

[2] 张建中 . 治理假新闻："后真相"时代欧洲国家的创新与实践 [J]. 新闻界，2017（06）：95–101.

[3] GILL S. Why Combatting Fake News Requires People And Technology—Working Together[EB/OL].（2019–09–15）. https：//www. knightfoundation. org/articles/why–combatting–fake–news–requires–people–and–technology–working–together.

[4] 华商网 . 人工智能是虚假新闻的"克星"[EB/OL].（2017–03–23）https：//www.sohu.com/a/129892845_119659..

[5] 蒋梦婷 . 虚假新闻检测技术的应用 [J]. 网络安全技术与应用，2021（04）：54–55.

[6] GuoBin，DingYasan，YaoLina，etal. Thefutureoffalse in formationdetectiononsocialmedia：Newperspectivesand trends[J]. ACM ComputingSurveys，2020（4）：68.

[7] CastilloC，MendozaM，PobleteB. Informationcredibilityon Twitter[C]//ProcoftheWebConf2020. NewYork：ACM，2011：675–684

[8] QazvinianV，RosengrenE，RadevD，etal. Rumorhasit：Identifyingmisinformationin microblogs[C]//Procofthe 2011 Conf on Empirical Methods in Natural Language Processing. Stroudsburg，PA：ACL，2011：1589–1599

[9] Pérez G RosasV，KleinbergB，Lefevre A，etal. Automatic detectionoffakenews [EB/OL].（2017–08–23）[2020–10–08]. https：//arxiv. org/pdf/1708. 07104. pdf .

[10] MaJing，Gao Wei，MitraP，etal. Detectingrumorsfrom microblogswithrecurrentneuralnetworks[C]//Procofthe 25thIntJointConfon ArtificialIntelligence. MenloPark，CA：AAAI，2016：3818–3824.

[11] 刘波，李洋，孟青，等. 社交媒体内容可信性分析与评价 [J]. 计算机研究与发展，2019（9）：19391952.

[12] JinZhiwei，CaoJuan，JiangYugang，etal. Newscredibility evaluation on microblog with a hierarchical propagation model[C]//Procofthe2014IEEEIntConfonDataMining. Piscataway，NJ：IEEE，2014：230–239.

[13] Jin Zhiwei，Cao Juan，Zhang Yongdong，et al. News certification by exploiting conflicting socialviewpointsin microblogs[C]//Procofthe30th AAAIConfon Artificial Intelligence. PaloAlto，CA：AAAI，2016：2972–2978.

[14]ShuKai，WangSuhang，Liu Huan. Beyondnewscontents：Theroleofsocialcontextforfakenewsdetection[C]//Procofthe20thACMIntConfon WebSearchandDataMining. NewYork：ACM，2019：312–320.

[15]MaJing，GaoWei，WongKF. RumordetectiononTwitter withtree G structuredrecursiveneuralnetworks[C]//Procof the 56th Annual Meeting of the Association for ComputationalLinguistics. Stroudsburg，PA：ACL，2018：1980–1989.

[16] 虎嵩林，赵军，唐杰，秦兵，石川，颜水成 . 前言 [J]. 计算机研究与发展，2021，58（07）：1351–1352.

[17]NobleCC，CookDJ. Graph G basedanomalydetection[C]//Procofthe 9th ACM SIGKDD Int Confon Knowledge DiscoveryandData Mining. New York：ACM，2003：631，636.

[18] SeoJ，Mendelevitch O. Identifyingfraudsandanomaliesin Medicare G Bdataset[C]//Procofthe39thAnnualIntConfof theIEEE Engineering in Medicine and Biology Society. Piscataway，NJ：IEEE，2017：3664–3667.

[19] ColladonAF，RemondiE. Usingsocialnetworkanalysisto prevent money laundering [J]. Expert Systems with Applications，2017，67：4958 1451 .

[20]ManjunathaH C，Mohanasundaram R. BRNADS：Bigdata real G timenodeanomalydetectioninsocialnetworks [C]// Procofthe2ndIntConfonInventiveSystemsandControl. Piscataway，NJ：IEEE，2018：929–932.

[21]SánchezPI，MüllerE，LaforetF，etal. Statisticalselection ofcongruentsubspacesforminingattributedgraphs[C]// Procofthe13thIEEEIntConfonDataMining. Piscataway，NJ：IEEE，2013：647–656 .

[22] SánchezP I，Müller E，Irmler O，etal. Localcontext selectionforoutlierrankingingraphswithmultiplenumeric nodeattributes[C]//Procofthe26thIntConfonScientific and Statistical Database Management. Piscataway，NJ：IEEE，2014：16.

[23] PerozziB，AkogluL，SánchezPI，etal. Focusedclustering andoutlierdetectioninlargeattributedgraphs[C]//Procofthe20th ACM SIGKDDIntConfon KnowledgeDiscovery andDataMining. NewYork：ACM，2014：1346–1355 .

[24]DaiHanbo，ZhuFeida，Lim EP，etal. Detectinganomalies inbipartitegraphswithmutualdependencyprinciples[C]// ProcoftheIEEE12thIntConfonDataMining. Piscataway，NJ：IEEE，2012：171–180.

[25] TsangS，Koh Y S，Dobbie G，etal. SPAN：Finding co

llaborativefraudsinonlineauctions[J]. Knowledge G Based Systems，2014，71：389−408.

[26] ShehnepoorS，SalehiM，FarahbakhshR，etal. Netspam：A network G based spam detection framework for reviewsin onlinesocialmedia[J]. IEEE TransactionsonInformation ForensicsandSecurity，2017（7）：1585−1595 .

[27] CarvalhoLF M，TeixeiraCHC，MeiraW，etal. Provider G consumeranomalydetectionforhealthcaresystems [C]// Procofthe IEEE Int Conf on Healthcare Informatics. Piscataway，NJ：IEEE，2017：229−238.

[28] Ranshous S，Shen Shitian，Koutra D，et al. Anomaly detection in dynamic networks：A survey [J]. Wiley InterdisciplinaryReviews：ComputationalStatistics，2015，7（3）：223−247.

[29] ManzoorE，MilajerdiSM，AkogluL. Fastmemory G efficient anomalydetectioninstreamingheterogeneousgraphs[C]// Procofthe22nd ACM SIGKDDIntConfon Knowledge DiscoveryandDataMining. New York：ACM，2016：1035 −1044 .

[30] 陈波冯，李靖东，卢兴见，沙朝锋，王晓玲，张吉 . 基于深度学习的图异常检测技术综述 [J]. 计算机研究与发展，2021（7）：1436−1455.

[31] WuZhizheng，EvansN，Kinnunen T，etal. Spoofingand countermeasuresfor speaker verification：A survey [J]. SpeechCommunication，2015：130−153.

[32] 王成龙，易江燕，陶建华，马浩鑫，田正坤，傅睿博 . 基于全局 − 时频注意力网络的语音伪造检测 [J]. 计算机研究与发展，2021（7）：1466−1475.

[33] 亓鹏，曹娟，盛强 . 语义增强的多模态虚假新闻检测 [J]. 计算机研究与发展，2021（7）：1456−1465.

[34]Amrita，Bhattacharjee，舒凯，高旻，刘欢 . 网络信息生态系统中的虚假信息：检测、缓解与挑战 [J]. 计算机研究与发展，2021（7）：1353−1365.

[35] 吴果中，李泰儒 . 用区块链技术打击虚假新闻——Userfeeds 与 PressCoin 模式介绍 [J]. 新闻战线，2018（13）：88−90.

作者简介：李净（1983−），女，河南南阳，《中国传媒科技》杂志社编辑部主任，研究方向：新闻传播。

（责任编辑：陈旭管）

